8B

УДК 004.048:004.622

Т.А. Васяева, Ю.А. Скобцов
Донецкий национальный технический университет, г. Донецк, Украина
vasyaeva_tanya@tr.dn.ua, skobtsov@kita.dgtu.donetsk.ua,

Эволюционный подход к формированию знаний для медицинских экспертных систем с учетом неопределенности данных

Разработан аппарат генетического программирования для прогнозирования СВСГД. Предложен метод получения продукционных правил для прогнозирования высокой степени риска СВСГД в условиях неопределенности некоторых параметров. Проведены исследования и приведены результаты использования методов на реальных медицинских данных.

Формирование базы знаний является одной из наиболее трудоемких задач при разработке экспертных систем (ЭС). Один из подходов формирования знаний заключается в разработке программ, способных обучаться под руководством эксперта-учителя. При этом учитель предъявляет программе примеры реализации некоторого концепта, а задача программы состоит в том, чтобы извлечь из предъявленных примеров набор атрибутов и значений, определяющих этот концепт. Данная работа является развитием [1], где для извлечения знаний в виде системы продукций используется аппарат генетического программирования. В отличие от предыдущих работ, где фактически используется двоичная логика, в настоящей работе применяется троичная логика, которая позволяет учитывать неопределенность (или отсутствие) значений некоторых параметров как на этапе обучения, так и в процессе эксплуатации ЭС.

Неопределенность данных

При работе с медицинскими данными, достаточно часто возникает ситуация, когда некоторые параметры неизвестны. Это затрудняет как и обучение системы, так и ее тестирование, а также использование. При формировании обучающих данных используются данные, предоставленные медицинскими работниками. Как правило, эти данные собираются по карточкам пациентов, которые находились на лечении несколько лет назад. Поэтому при отсутствии некоторой информации практически невозможно ее восстановить. Классические автоматизированные методы формирования знаний на базе машинного обучения (machine learning) работают, если известны все выделенные факторы риска для каждого пациента. Поэтому, если какой-нибудь параметр неизвестен только у одного пациента, необходимо либо удалить пациента из обучающей выборки, либо удалить данный параметр из списка факторов риска. Так как в большинстве случаев у разных пациентов отсутствуют данные о различных факторах риска, формирование обучающей выборки в этом случае выполняется с существенной потерей данных.

После разработки системы список входных параметров, как правило, уже определен и для корректной работы системы все информативные составляющие должны быть заполнены. При тестировании отсутствие информации сказывается на достоверности результата или невозможности диагностирования в целом.

Целью проектируемой системы в данной статье является получение продукционных правил для диагностирования заболевания в условиях неопределенности некоторых входных данных (на примере определения высокой степени риска синдрома внезапной смерти грудных детей – (СВСГД) – одного из малоизученных и загадочных заболеваний).

В данной задаче в качестве обучающего множества используются реальные данные обследования 240 пациентов (120 детей, которые умерли в Донецкой области от СВСГД, и контрольная группа из 120 живых детей на первом году жизни). Данные составляют информацию общего характера и образа жизни беременных, а также перенесенные заболевания и результаты некоторых анализов.

Генетическое программирование

Для решения поставленной задачи предложено использовать генетическое программирование (ГП) [2]. Решение задачи на основе ГП можно представить следующей последовательностью действий.

1. Установка параметров эволюции;
2. Инициализация начальной популяции;
3. Т:=0;
4. Оценка особей, входящих в популяцию;
5. Т:=Т+1;
6. Отбор родителей;
7. Создание потомков выбранных пар родителей – выполнение оператора кроссинговера;
8. Мутация новых особей;
9. Расширение популяции новыми порожденными особями;
10. Сокращение расширенной популяции до исходного размера;
11. Если критерий остановки алгоритма выполнен, то выбор лучшей особи в конечной популяции – результат работы алгоритма. Иначе переход на шаг 4.

Предлагается следующий метод кодирования особей для генетического программирования. Каждая особь представляет собой дерево, которое соответствует синтаксическому выражению, представляющее множество правил в дизъюнктивной нормальной форме.

На рис. 1 представлен пример дерева в дизъюнктивной нормальной форме. Дерево представлено 3-мя правилами. Такое представление особи значительно упрощает интерпретацию результата. В данном примере расшифровка будет следующей:

ЕСЛИ правило 1 ИЛИ правило 2 ИЛИ правило 3, ТО результат 1, ИНАЧЕ результат 2.

Популяция особей (потенциальных решений) состоит из набора деревьев, сгенерированных случайным образом. Генерация каждого дерева, как описано ниже, происходит рекурсивно, начиная с первого функционального узла ИЛИ и его аргументов. По построенному специальным образом дереву можно получить систему продукций, которая классифицирует с заданной точностью данные обучающей выборки.

Входное обучающее множество должно быть представлено в виде булевых переменных. Для этого исходные данные были преобразованы следующим образом:

- место жительства (город – 1, село – 0);
- возраст матери на момент родов (полных лет) <17;
- возраст матери на момент родов (полных лет) <25;

– возраст матери на момент родов (полных лет) <30;
– возраст матери на момент родов (полных лет) >31;
– место работы матери, профвредность (да – 0 , нет – 1);
– и др.

Терминальное множество состоит из факторов риска, которые после предобработки представляют собой булевые переменные и соответствуют листьям дерева. Функциональное множество состоит из логических операций: AND, OR, NOT, которые представляют внутренние вершины дерева.

В качестве фитнесс-функции рассматривается: доля пациентов с правильно поставленным диагнозом. Переменная диагноза принимает булевые значения 0 или 1. Единица соответствует положительному диагнозу (высокой степени риска СВСГД) и ноль отрицательному (низкой степени риска СВСГД). Значение фитнесс-функции для особей с правильным диагнозом принимает значение 1, а для особей с неправильным диагнозом принимает значение 0.

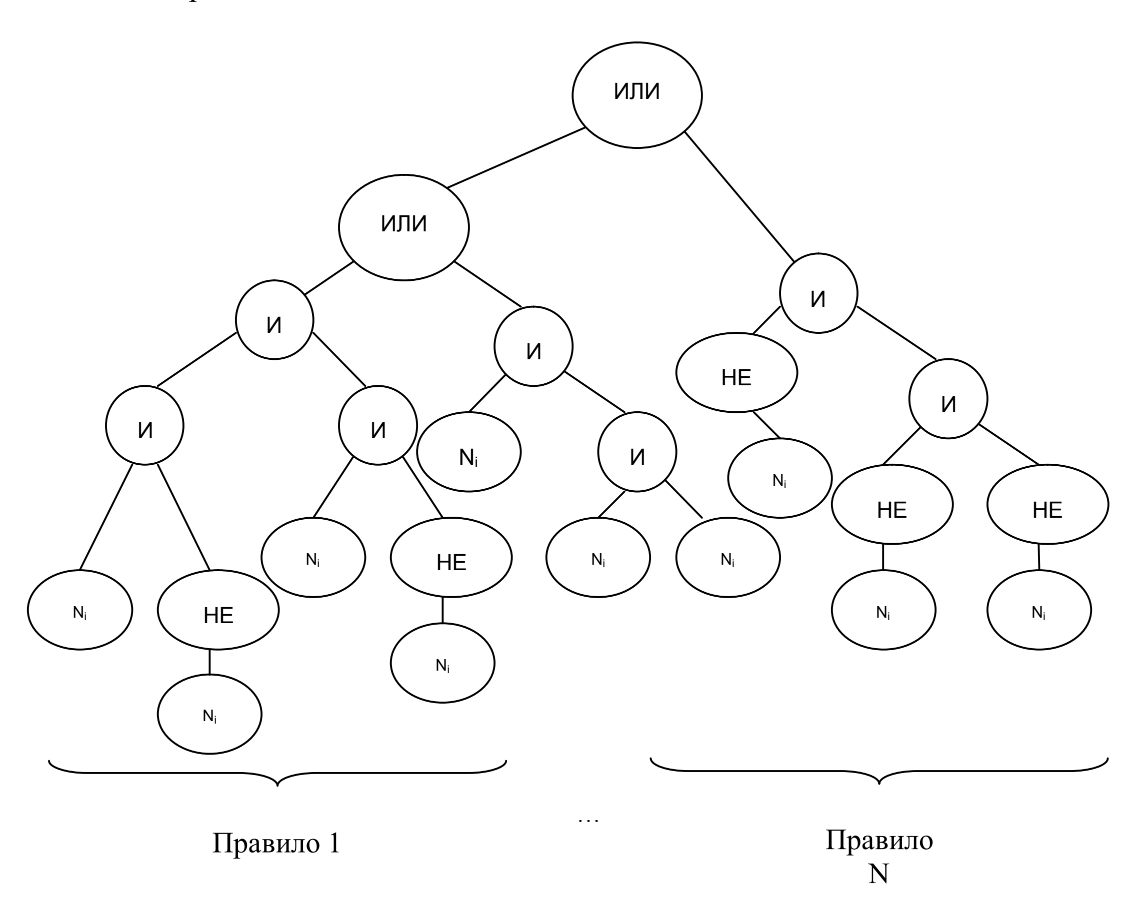


Рисунок 1 – Пример дерева в дизъюнктивной нормальной форме

С целью минимизации потери данных при обучении и расширения возможностей диагностирования при неизвестных значениях некоторых факторов риска предлагается использовать троичную логику. При этом переменные могут принимать три логические значения {0,1,*}, где ‘*’ представляет неопределенное значение (это 0 или 1, но неизвестно, что именно). Подобный подход применяется во многих отраслях науки и техники, например, при проектировании цифровых систем с использованием логического моделирования в троичной (или многозначной) логике [3].

В табл. 1 – 3 приведены таблицы истинности для следующих логических функций: И, ИЛИ и НЕ.

Таблица 1

N_1	N_2	И
0	0	0
0	1	0
1	0	0
1	1	1
*	0	0
*	1	*
*	*	*

Таблица 2

N_1	N_2	ИЛИ
0	0	0
0	1	1
1	0	1
1	1	1
*	0	*
*	1	1
*	*	*

Таблица 3

N_1	НЕ
0	1
1	0
*	*

Применение системы, которая оперирует с неизвестными состояниями, позволит выполнять диагностику даже при отсутствии некоторых параметров, что не приведет к невозможности функционирования разработанной системы. На этапе обучения такой подход позволит сформировать оптимально полный набор входных параметров и не упустить важные, информативные параметры.

Генерация начальной популяции

На данном этапе происходит генерация начальной популяции, в соответствии с заданными параметрами. Популяция состоит из набора деревьев, сгенерированных случайным образом. Генерация каждого дерева происходит рекурсивно, начиная с генерации первым функционального узла ИЛИ и его аргументов. В качестве аргументов на первом шаге может быть только узел ИЛИ. Далее для каждого дочернего узла случайным образом определяется тип и значения его аргументов по следующим принципам:

– после узла ИЛИ может быть только функциональный узел (значениями которого могут быть – ИЛИ или И);
– после узла И может быть функциональный узел (значениями которого могут быть – И или НЕ) или терминальные узлы;
– после узла НЕ может быть только терминальный узел.

Процесс выполняется по левой ветви до тех пор, пока не будет выбран дочерним терминальный узел. Затем генерируются правые ветви.

Вероятность функционального и терминального узлов меняется по следующему принципу: чем ниже вершина, тем больше вероятность терминального узла и меньше функционального. Для функционального узла на каждом последующем шаге увеличивается вероятность узла И и уменьшается вероятность узла ИЛИ.

При формировании дерева в одной ветви ИЛИ (т.е. для одного правила) не используется один и тот же терминальный символ более одного раза.

Предусмотрены методы создания деревьев: полный, растущий и комбинированный.

Отбор родителей. Предложено использовать отбор пропорционально значению целевой функции, реализованный методом рулетки или турниром. При этом если два или более потомка имеют одинаковую фитнесс-функцию, то выбирается дерево минимальной сложности.

Кроссинговер

Для древообразной формы представления используются следующие три основных оператора кроссинговера:

- узловой кроссинговер;
- кроссинговер поддеревьев;
- смешанный.

Учитывая строго определенное представление дерева необходимо модифицировать операторы кроссинговера.

В узловом операторе кроссинговера обмен возможен только для терминальных узлов.

В кроссинговере поддеревьев родители могут обмениваться только поддеревьями ветви И.

При смешанном операторе кроссинговера для некоторых узлов выполняется узловой оператор кроссинговера, а для других – кроссинговер поддеревьев.

Также предлагается выполнять оператор кроссинговера для худшего правила в дереве. Правило считается худшим, у которого целевая функция имеет минимальное значение. Каждое правило можно рассматривать как отдельное дерево, способное решать поставленную задачу, поэтому вычисление фитнесс-функции для каждого правила в отдельности логически обосновано.

Вычисление фитнесс-функции не только для каждого правила в отдельности, но и каждого узла И также имеет смысл. При выполнении оператора кроссинговера поддеревьев предлагается осуществлять поиск точки разрыва следующим образом: вычисляется фитнесс-функция для каждого узла И начиная с первого снизу. Если значение фитнесс-функции для узла И, находящегося выше, хуже, чем на предыдущем шаге, то обмену подлежит один из узлов аргументов данного узла И.

Мутация

Для деревьев используются следующие операторы мутации:

- узловая;
- усекающая;
- растущая.

Как и в случае с оператором кроссинговера оператор мутации должен быть модифицирован.

Узловая мутация выполняется для терминального узла или первой снизу вершины ИЛИ.

Усекающая мутация выполняется только для узлов И или НЕ.

При растущей мутация ветви наращиваются согласно правилам инициализации деревьев.

Сокращение дерева

Предлагается использовать оператор сокращения дерева. Как и оператор кроссинговера или мутации, данный оператор выполняется с определенной вероятностью. Если количество правил в дереве превышает определенный порог, то обрезается целое правило. Если количество правил не превышает указанное число, то обрезается худшая часть дерева в худшем правиле, т.е. выполняется усекающая мутация.

Редукция

Предлагается использовать выполнения следующих вариантов редукции:

– элитная стратегия;
– чистая замена;
– равномерная случайная замена (с указанием количества заменяемых особей в %).

При тестировании на реальных медицинских данных получили следующие результаты. На рис. 2 представлены результаты экспериментов: зависимость правильной классификации от количества неизвестных состояний на входах в %. На рис. 3 представлены результаты экспериментов: зависимость нераспознанных диагнозов от количества неизвестных состояний на входах в %.

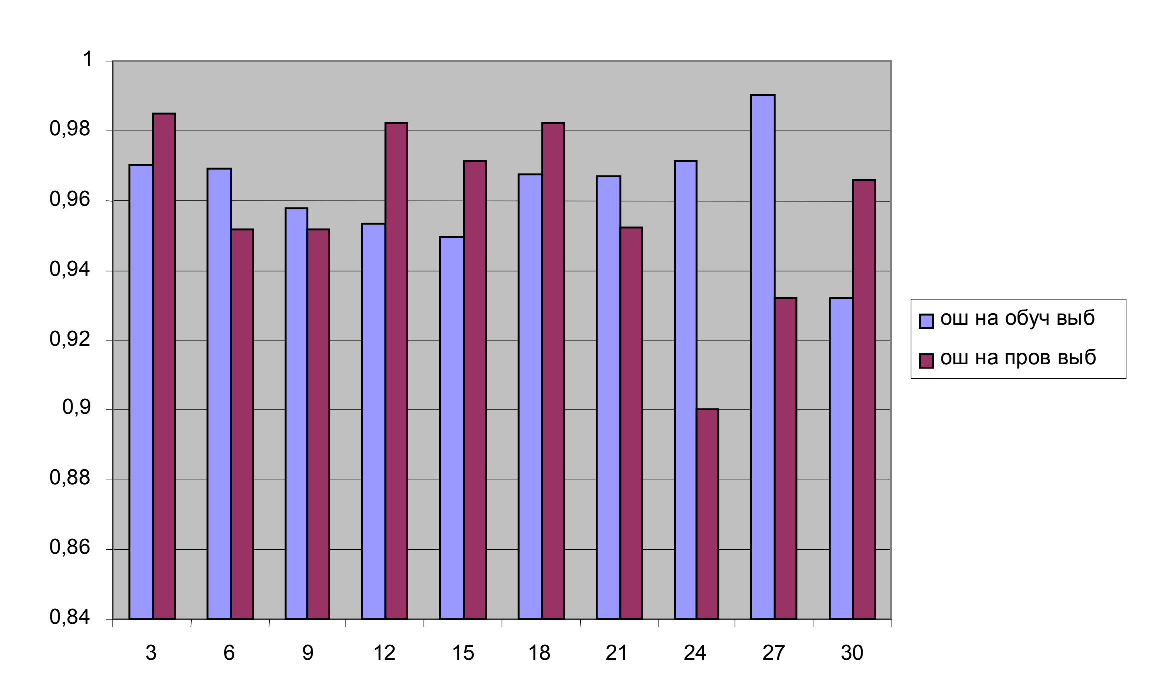


Рисунок 2 – Зависимость правильной классификации от количества неизвестных состояний на входах в %

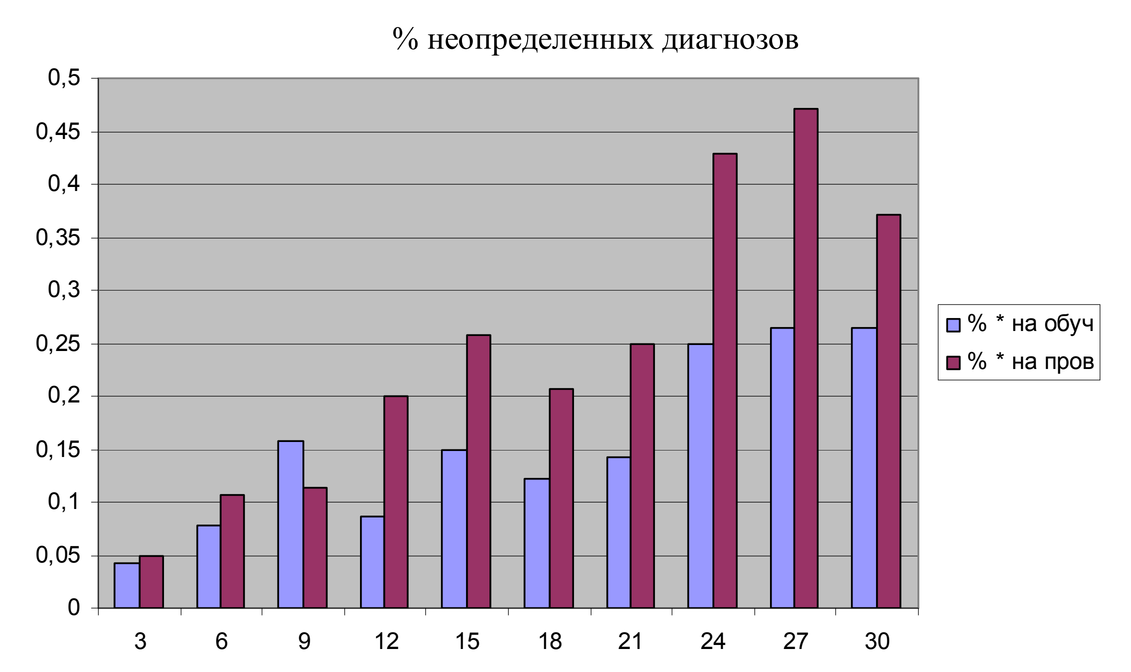


Рисунок 3 – Зависимость нераспознанных диагнозов от количества неизвестных состояний на входах в %

Выводы

Таким образом, получил дальнейшее развитие метод прогнозирования на основе генетического программирования, что позволило получить продукционные правила для прогнозирования высокой степени риска СВСГД в условиях неопределенности некоторых параметров. Предложенный метод протестирован на примере прогнозирования СВСГД, но может быть использован и при решении других задач медицинской диагностики и прогнозирования.

Литература

1. Васяева Т.А., Скобцов Ю.А. Разработка экспертных систем медицинской диагностики с явным представлением продукционных правил на основе ГП // Тези міжнародної наукової конференції «Інтелектуальні системи прийняття рішень і проблеми обчислювального інтелекту (ISDMCI’2008)». – Т. 3, Ч. 1. – Херсон: ХНТУ, 2008. – 192 с.
2. Скобцов Ю.А. Основы эволюционных вычислений: Навчальний посібник. – Донецьк: ДонНТУ, 2008. – 326 с.
3. Скобцов Ю.А., Скобцов В.Ю. Логическое моделирование и тестирование цифровых устройств. – Донецк: ИПММ НАНУ, ДонНТУ, 2005. – 436 с.

Т.О. Васяєва, Ю.О. Скобцов

Еволюційний підхід до формування знань для медичних експертних систем з урахуванням нез’ясованості даних

Розроблено апарат генетичного програмування для прогнозування СРСН. Запропоновано метод отримання продукційних правил для прогнозування високого ступеня ризику СРСН при деяких нез’ясованих параметрах. Виконано експерименти та наведено результати використання методу на реальних медичних даних.

Статья поступила в редакцию 17.07.2008.